

EgoMind: Activating Spatial Cognition through Linguistic Reasoning in MLLMs

Supplementary Material

A. Implementation Details

A.1. Training Strategy

Our training pipeline consists of two stages: Supervised Fine-Tuning (SFT) for initializing spatial reasoning ability and aligning the model with the EgoMind CoT format, followed by Reinforcement Learning (RL) to further enhance structured reasoning quality through GRPO.

Supervised Fine-Tuning. Based on the 5K automatically generated SFT samples described in Fig. 3 of the main paper, we fine-tune Qwen2.5-VL-7B using the LLaMA-Factory framework to provide the model with initial spatial reasoning ability and align its outputs with the EgoMind CoT format, forming a strong foundation for the subsequent RL stage. During SFT, 16 frames are uniformly sampled from each video, and the maximum pixel budget is constrained to 200,704 ($256 \times 28 \times 28$). Training is conducted for 3 epochs with a learning rate of 5×10^{-6} using a cosine decay schedule and a warmup ratio of 0.1. We employ the AdamW optimizer and train in `bf16` precision to improve memory efficiency and stability. The LLM and projector components are set to be trainable, while the ViT backbone remains frozen.

To ensure consistency with the RL stage, we append a structured instruction prompt to each question, guiding the model to produce outputs that adhere to the EgoMind CoT format:

Task Prompt

You should first think about the reasoning process in the mind and then provide the user with the answer. The reasoning process and answer are enclosed within `<think>``</think>` and `<answer>``</answer>` tags, respectively, i.e., `<think>`reasoning process here `</think>``<answer>`answer here `</answer>`

Reinforcement Learning. During the RL phase, we train the MLLM on 20K samples using the GRPO algorithm implemented in the EasyR1 framework. We set the batch size to 64, the learning rate to 1×10^{-6} , and apply a weight decay of 1.0×10^{-2} . The AdamW optimizer is adopted with `bf16` precision. To balance effective policy updates with controlled divergence from the reference model, we use a KL penalty coefficient of 1×10^{-4} . For each prompt, the policy generates 8 candidate reasoning paths to compute group-wise rewards, using a temperature of 1.0 and top- p of

0.99. The maximum response length is capped at 2048 tokens.

The total reward is defined as a weighted sum of a format reward (R_{format}) and an accuracy reward (R_{accuracy}), with weights 0.2 and 0.8, respectively. The format reward is binary: $R_{\text{format}} = 1$ if the model output strictly adheres to the required `think-answer` structure, and 0 otherwise. The accuracy reward R_{accuracy} evaluates the content contained within the `<answer>` and `</answer>` tags. For multiple-choice questions, we assign a discrete score of 0 or 1 based on exact matching of the predicted option (A/B/C/D). For numerical questions, we compute accuracy using the Mean Relative Accuracy (MRA) metric, which measures the relative closeness between the predicted value and the ground truth.

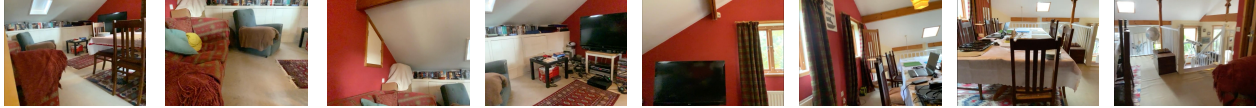
A.2. Inference Strategy

To ensure fair comparisons across models of varying architectures, we strictly standardize our evaluation protocol. For closed-source models (e.g., GPT-4.1, GPT-5, and Gemini 2.5 Pro), we apply the identical CoT prompt detailed in Section A.1. For open-source MLLMs, we report the higher score between direct generation and CoT prompting to avoid penalizing models with weaker instruction-following capabilities. Furthermore, we evaluate all reasoning models under their official default configurations (e.g., “dynamic thinking” for Gemini 2.5 Pro and “medium reasoning effort” for GPT-5).

Regarding visual inputs, we constrain the maximum image resolution to $256 \times 28 \times 28$ across all benchmarks to maintain strict consistency with our training phase. Additionally, for video-based tasks specifically, we uniformly sample 16 frames per video sequence.

A.3. Data Construction

Supervised Fine-Tuning Data For the SFT stage, our goal is to construct a compact yet diverse dataset that enables MLLMs to learn the EgoMind CoT format under strict cost constraints. To achieve this, we sample approximately 5K instances from the SpaceR-91K dataset. During sampling, we filter out trivial or overly ambiguous cases and enforce a more uniform distribution across different question types and answer patterns to reduce dataset-induced bias. The corresponding EgoMind CoT annotations are automatically generated using the pipeline described in Sec. 3.4 of the main paper. Furthermore, to guarantee the reliability of the generated data, we employ Gemini 2.5 Pro driven



Task

I am standing on the red carpet, facing the white cabinet with books, so where are the laptops located relative to me?
 A. left B. right C. front D. back

EgoMind-7B

<think>

To solve the problem, I will first establish the observer's position and orientation as described in the question. Then, by analyzing the visual information from the video frames, I will locate the white cabinet with books and the laptops. Finally, I will determine the spatial relationship between the observer and the laptops to answer the question.

<ROLE_PLAY_CAPTION>

I am standing in a bright room. As I take two steps forward, I notice a red-patterned sofa to my left, and directly in front of me is a green armchair placed at an angle. Behind it, I see a built-in cabinet with books on its top surface. As I turn my view approximately 90 degrees to the right, there is a television with audio equipment underneath. To the left of the TV, a small, low black table is cluttered with items. Continuing to pan my view another 90 degrees to the right, I see a dining table with a white tablecloth and wooden chairs, which has two laptops on it. Further to the right in the background, a staircase with a white railing is visible.

</ROLE_PLAY_CAPTION>

<SPATIAL>

1. **Task-relevant objects:**

- * [OBSERVER]: The person asking the question.
- * [WHITE-CABINET-0]: A white, low cabinet against a red wall, holding books.
- * [LAPTOP-0], [LAPTOP-1]: Two laptops seen on a table.

2. **Spatially proximate objects:**

- * [DINING-TABLE-0]: A table with a white tablecloth.
- * [TV-0]: A large TV on a stand.

3. **Spatial relationships:**

- * The [OBSERVER] is standing on [RED-CARPET-0].
- * The [OBSERVER] is facing [WHITE-CABINET-0].
- * [TV-0] is to the right of the [OBSERVER].
- * [RED-CARPET-0] is between [WHITE-CABINET-0] and [DINING-TABLE-0].
- * [LAPTOP-0] and [LAPTOP-1] are located on [DINING-TABLE-0].

</SPATIAL>

1. The question establishes that the observer is standing on the [RED-CARPET-0] and facing the [WHITE-CABINET-0].
2. Since [RED-CARPET-0] is between [WHITE-CABINET-0] and [DINING-TABLE-0], [DINING-TABLE-0] is located directly behind the observer's initial position.
3. Since [LAPTOP-0] and [LAPTOP-1] are on [DINING-TABLE-0], the laptops are behind the observer.

</think>

<answers> D. back </answers>

Qwen2.5VL-7B

B. right

Figure D. A case study of relational reasoning with the Qwen2.5-VL-7B model enhanced by the EgoMind framework.

by specifically tailored prompts to conduct comprehensive quality checks and filtering:

- **Hallucination Check:** We verify whether the finally merged chain-of-thought content factually conflicts with the input video frames, ensuring that no erroneous information is introduced during the Merge stage.
- **Logical Consistency:** We strictly examine the consistency between the PSA/RPC context and the Reasoning section. This guarantees that the reasoning conclusions are logically derived from the evidence provided by PSA and RPC.
- **Format & Correctness:** We check whether the final extracted answer is correct by comparing it against the ground truth labels, and we ensure that the output format strictly complies with the training requirements.

Reinforcement Learning Data For the RL stage, we further sample 20K instances from the SpaceR-91K dataset. Since RL relies solely on outcome-based rewards and does not require CoT supervision, we remove extreme cases that are excessively easy or unsolvable. From the remaining pool, we select 20K moderately challenging samples to provide sufficient difficulty for policy improvement while maintaining stable reward signals.

A.4. Benchmarks

To comprehensively evaluate the spatial reasoning capability of EgoMind, we consider four representative benchmarks that cover a diverse set of spatial perception and reasoning tasks.

VSI-Bench assesses an MLLM's ability to perceive, memorize, and reason about physical spaces through continuous visual observation. It contains 5,000 QA pairs across 288 real-world indoor videos sourced from ScanNet, ScanNet++, and ARKitScenes, and includes tasks such as object counting, distance estimation, relative direction prediction, and route planning.

SPAR-Bench provides over 7,200 human-verified QA samples spanning 20 spatial reasoning tasks, ranging from basic geometric perception to high-level relational reasoning. It uniquely employs only static images (single-view or multi-view), enabling a pure evaluation of a model's ability to infer 3D spatial structure from discrete viewpoints without temporal information.

SITE-Bench integrates 30 existing spatial-intelligence datasets and augments them with newly designed tasks, offering a unified multiple-choice framework for systematic evaluation of spatial reasoning in MLLMs. In our experiments, we adopt the video-based subset of SITE-Bench, which contains 3,808 video QA tasks covering diverse spa-

Table E. Performance evaluation of Qwen2.5-VL models at different scales (3B and 7B) on the VSIBench benchmark. The table details the step-wise performance gains, comparing the base models with their counterparts enhanced through Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL).

Models	Params	Version	Overall	Numerical Question				Multiple-Choice Question			
				Obj. Cnt.	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
Qwen2.5-VL-3B	3B	Base	27.68	27.06	19.78	25.37	19.24	32.54	39.26	28.87	22.33
		+SFT	37.08	36.48	21.74	36.88	30.66	37.04	45.14	22.68	53.56
		+RL	45.44	44.96	31.29	59.07	37.60	41.69	47.52	23.20	55.66
Qwen2.5-VL-7B	7B	Base	30.02	23.75	10.50	36.98	35.31	38.45	37.09	28.87	28.48
		+SFT	42.33	43.20	23.87	50.34	29.03	41.41	47.42	29.38	57.44
		+RL	50.16	54.51	37.94	67.12	40.35	44.08	47.21	31.96	58.41

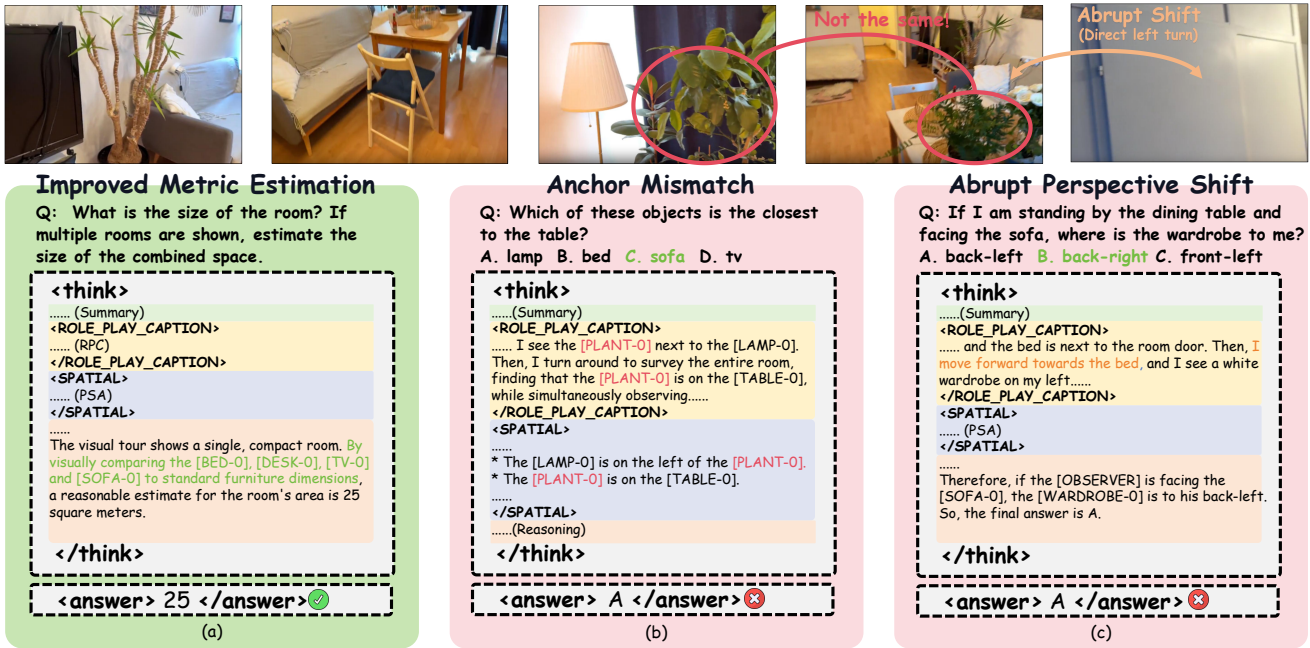


Figure E. Case studies of EgoMind.

tial understanding scenarios.

SPBench comprises 1,328 QA pairs divided into two subsets: SPBench-SI (single-image, 1,009 QA) and SPBench-MV (multi-view, 319 QA). It is specifically designed to measure a model’s geometric understanding, object enumeration ability, and multi-view spatial synthesis. All samples are derived from indoor scenes in the ScanNet dataset.

B. Extended Ablation Studies

B.1. Generalization Across Model Scales

To evaluate the generalization ability of the EgoMind CoT across different MLLM scales, we fine-tune Qwen2.5-VL-3B using the proposed framework and assess its performance on VSI-Bench. The results are presented in Table E.

As shown in Table E, EgoMind yields consistent im-

provements on both Qwen2.5-VL-3B and Qwen2.5-VL-7B. With only the SFT stage, performance increases from 27.68 → 37.08 on the 3B backbone and from 30.02 → 42.33 on the 7B backbone, indicating that the EgoMind CoT effectively equips MLLMs with structured spatial reasoning abilities. When reinforcement learning is further introduced, the scores improve substantially to 45.44 and 50.16 for the 3B and 7B models, respectively, demonstrating the strong synergy between CoT-based supervision and RL-driven refinement.

It is worth noting that EgoMind requires only 5K automatically generated CoT samples for SFT and 20K QA-only samples for RL—significantly fewer than competing methods. These results highlight the high data efficiency of EgoMind and its ability to activate robust spatial cognition purely through carefully designed linguistic reasoning,



Task

If I am standing by the bed and facing the dressing table, is the hanger to my front-left, front-right, back-left, or back-right?
 A. back-right B. front-left C. **back-left** D. front-right

Gemini-2.5-Pro (Enhanced)

<think>

My goal is to determine the position of the hanger relative to an observer who is standing by the bed and facing the dressing table. First, I will analyze the video frames to build a mental map of the bedroom, identifying the key objects: the bed, the dressing table, and the hanger. Then, I will establish a plausible position and orientation for the observer based on the problem description. Finally, using this established viewpoint, I will determine the relative location of the hanger and select the correct option from the choices provided.

<ROLE_PLAY_CAPTION>

I'm peeking into a bedroom from a wooden doorway. The floor is hardwood. In the center of the room is a large bed with a light-colored patterned duvet and a wooden headboard. To the left of the bed, there's a small gold-colored metal and glass nightstand with a lamp on it. To the right of the bed, there's a patterned rug on the floor and a black mesh trash can. Against the far wall, to the right of the bed, stands a wooden chest of drawers. Between the head of the bed and the chest of drawers, a red jacket is hanging on a wall hanger.

I step further into the room, moving along the right side of the bed. Above the bed's headboard is a large, rectangular mirror with a silver-colored frame. I pan left, getting a full view of the bed and the mirror. The reflection shows me holding the camera.

Now I move forward, towards the corner of the room, passing the right side of the bed. The wooden chest of drawers is now directly to my left. I pan right, past the chest of drawers. There's a wooden chair with a light-colored seat cushion next to it. Beyond the chair is a window with green, yellow, and blue patterned curtains tied back. The window has a wide wooden sill.

Continuing to pan right, past the window, I see the room opens into a corner nook. I turn into this nook. On the far wall of the nook is a white dressing table with several drawers and a collection of cosmetics and a mirror on its surface. A black and chrome swivel chair is tucked under it. To the right of the chair, a red hairdryer is mounted on a stand. White double doors for a closet are visible on the right wall of this nook. I'm now standing directly in front of the dressing table.

</ROLE_PLAY_CAPTION>

<SPATIAL>

1. **Task-Relevant Objects:**

- * [OBSERVER] The person standing by the bed and facing the dressing table.
- * [BED-0] The large bed in the main area of the room.
- * [DRESSING-TABLE-0] The white vanity located in a nook to the right side of the room.
- * [HANGER-0] The wall hanger holding a red jacket.

2. **Surrounding Objects:**

- * [CHEST-OF-DRAWERS-0] A wooden chest of drawers against the back wall.
- * [MIRROR-0] A large rectangular mirror mounted on the wall above [BED-0].
- * [NIGHTSTAND-0] A glass and metal nightstand to the left of [BED-0].
- * [NIGHTSTAND-1] Another nightstand to the right of [BED-0].
- * [CHAIR-0] A wooden chair between [CHEST-OF-DRAWERS-0] and a window.
- * [WINDOW-0] The window on the right wall of the main room area.
- * [CHAIR-1] A black swivel chair at [DRESSING-TABLE-0].
- * [CLOSED-DOORS-0] White doors in the dressing table nook.

3. **Positional Relationships:**

- * The [OBSERVER] is standing by [BED-0], facing [DRESSING-TABLE-0]. A logical position for the [OBSERVER] is on the right side of [BED-0], as this provides a clear line of sight and path to [DRESSING-TABLE-0].
- * [BED-0] is in the center of the main part of the room.
- * [HANGER-0] is mounted on the back wall, between the head of [BED-0] and [CHEST-OF-DRAWERS-0].
- * [CHEST-OF-DRAWERS-0] is against the back wall, to the right of [BED-0] and [HANGER-0].
- * [DRESSING-TABLE-0] is located in a nook that is past the [WINDOW-0] on the right side of the room. To face it from the main room, one must look towards the right wall.

</SPATIAL>

- The problem requires me to assume the identity of an [OBSERVER] who is "standing by the bed" and "facing the dressing table".
- Based on the spatial analysis, [BED-0] is in the center of the room. [DRESSING-TABLE-0] is in a nook on the right side of the room. To face [DRESSING-TABLE-0] while being "by the bed", the [OBSERVER] would most logically be standing on the right side of [BED-0], looking towards the right wall where the dressing table nook is located.
- In this position, the [OBSERVER]'s front is directed toward [DRESSING-TABLE-0].
- The [OBSERVER]'s left side would be towards the head of [BED-0] and the back wall. The [OBSERVER]'s right side would be towards the foot of [BED-0] and the room entrance. The [OBSERVER]'s back would be towards the left side of the room.
- The [HANGER-0] is located on the back wall, near the head of [BED-0].
- Therefore, from the [OBSERVER]'s perspective (standing on the right of the bed, facing right), the [HANGER-0] is located behind them and to their left.
- This corresponds to the "back-left" quadrant relative to the [OBSERVER].

</think>

<answer> C. back-left </answer>

Gemini-2.5-Pro (Original)

A. back-right

Figure F. A case study where Gemini 2.5 Pro is guided by an EgoMind CoT prompt to solve a complex spatial relationship problem. This case illustrates that our proposed framework can be used as a zero-shot prompting strategy to unlock the spatial understanding and reasoning capabilities of powerful closed-source models.

without relying on additional multi-modal data or explicit 3D supervision.

B.2. Intermediate Results Verification

To rigorously verify the faithfulness and reliability of our generated reasoning chains, we employ Gemini 2.5 Pro as an independent judge to audit the intermediate reasoning traces on the VSI-Bench validation set.

We evaluate the intermediate results across two key dimensions: (i) **Visual Fidelity**, which measures whether the generated RPC and PSA context accurately reflects the raw video frames; and (ii) **Logical Consistency**, which assesses whether the final answer logically stems from the reasoning chain. Specifically, we design detailed evaluation prompts that instruct the judge to assign a binary score (0 or 1)

to each reasoning trace for both dimensions. These binary scores are then averaged to compute the final aggregate percentages. Our evaluation reveals that the EgoMind CoT achieves a high visual fidelity of **98.93%** for RPC and **91.60%** for PSA, alongside an impressive **96.69%** logical consistency. This strong alignment between the intermediate reasoning steps and the final answer confirms that EgoMind's performance gains arise from reliable, grounded spatial perception, effectively mitigating the risk of spurious correlations or shortcut learning.

C. Qualitative Analysis

To further investigate the qualitative improvements brought by EgoMind CoT, we evaluate the representative open-source MLLM, Qwen2.5-VL-7B, by comparing its re-

You will solve a problem by first providing your thought process, then the final answer.

Response Format

Your output MUST strictly follow the structure outlined below. Pay close attention to which sections use tags and which do not. The overall layout is:

<reason>

[A concise, 3-5 sentence summary of your plan]

<ROLE_PLAY_CAPTION>

[A first-person, real-time narration of the scene]

</ROLE_PLAY_CAPTION>

<SPATIAL>

[A structured, step-by-step analysis of objects and their positions]

</SPATIAL>

[A step-by-step logical deduction leading to the answer]

</reason>

<answer>

[The final, concise answer]

</answer>

Instructions for the Content Inside <reason> </reason>

You MUST include the following four components inside the <reason> block, in this exact order:

1. **Plan Summary (No Tags):**

- Start immediately after the <reason> tag.

- **Goal:** Briefly explain the steps to solve the problem (e.g., "I will first traverse the room to identify all X, then map their coordinates to count them."). Summarize your plan in 3-5 concise sentences.

2. **Role Play Caption (Using <ROLE_PLAY_CAPTION> tags):**

- **Goal:** Narrate from a first-person perspective (like a camera operator or robot) what you see and hear in real-time. Follow the chronological order of the video frames.

- **Requirements:**

- **Perspective:** Fully adopt the "Ego-centric" view. Narrate your physical movement (e.g., "I move forward 2 meters," "I pan 45 degrees left").

- **Visuals:** Describe the environment structure (e.g., "L-shaped room") and landmark objects. Mention colors and shapes clearly.

- **Continuity:** Dynamically narrate changes. If an object disappears due to occlusion or frame exit, mention it (e.g., "The brown table passes out of view on my left").

- **Object Permanence:** If you see an object again after turning, note it (e.g., "I see the same chair I passed earlier").

- **Example:** "I am in an office copy room. In front of me is a large, light-gray Canon multifunction copier. A notice is taped to its front. As I pan left 90 degrees, the copier moves out of frame and I see a white door about 2 meters away. I move forward towards the door, then turn right and look at the ceiling, finding a square air vent."

- **IMPORTANT:** Keep descriptions factual. Focus on **Spatial Geometry** (depth, occlusion, relative placement) rather than artistic flair. Ignore the specific question here; focus on a complete scene reconstruction.

3. **Spatial Analysis (Using <SPATIAL> tags):**

- **Goal:** Create a structured, "geometric" analysis of the scene to ground your reasoning.

- **Requirements:** Your output inside this section **MUST** follow these numbered steps in order:

1. **Task-Relevant Objects:** Identify and label all objects mentioned in or relevant to the task. Use '[OBJECT-TYPE-N]' (e.g., '[TABLE-0]').

- **Indexing:** Start N=0. Increment N for "distinct" instances.

- **De-duplication:** Be careful NOT to label the same object twice if it appears in multiple frames. Treat them as one unique entity.

- **Observer:** If applicable, label the camera/viewer as '[OBSERVER]'.

2. **Surrounding Context:** Identify nearby reference objects that help define scale or location (e.g., "The floor tiles," "The door frame"). Label them if necessary for clarity.

3. **Positional Relationships & Metrics:** Describe the "absolute position" (e.g., "North-east corner") and "relative position" (e.g., "[CHAIR-0] is -1 meter to the left of [TABLE-0]") of all labeled objects.

- **Estimation with Priors:** When estimating metrics (size, distance, area), "explicitly combine visual evidence with common-sense priors" (e.g., "Standard doorways are ~2m high," "Floor tiles are typically 60cm x 60cm," "This is a residential living room, likely 3 - 5m wide"). Use these priors as benchmarks to refine your estimates.

4. **Logical Reasoning (No Tags):**

- Place this section "after" the </SPATIAL> closing tag.

- **Goal:** Provide a step-by-step chain-of-thought deduction.

- **Requirements:**

- **Integration:** You **MUST** refer to the labels (e.g., '[TABLE-0]') defined in the <SPATIAL> section.

- **Verification:** If the task involves counting or estimation, cross-check your <ROLE_PLAY_CAPTION> (visual flow) with your <SPATIAL> list (deduplicated entities) to ensure accuracy.

- **Final Conclusion:** Deduce the final answer based "strictly" on the evidence above.

Instructions for the <answer> </answer> Section

- Provide **ONLY** the final, concise answer (e.g., a single number, a direction, or a short phrase). Do not add any extra explanation.

Your Assignment

{{content}}

Figure G. A zero-shot prompt to activate the spatial understanding and reasoning capabilities of powerful closed-source models.

sponses with and without EgoMind-style reasoning. We categorize our qualitative analysis into relational understanding, metric consistency, and typical failure modes.

Relational Reasoning Capabilities. As visualized in Fig. D, EgoMind CoT successfully activates robust spatial cognition in Qwen2.5-VL-7B for relative position and direction tasks. While the vanilla model struggles to maintain spatial awareness across multiple views, the enhanced model demonstrates the ability to construct a coherent, linguistically grounded spatial graph. It accurately identifies task-relevant objects across continuous frames and integrates these visual cues into a well-structured reasoning chain to deduce complex spatial relationships seamlessly.

Insights on Metric Consistency. Beyond qualitative relational reasoning, EgoMind excels at bridging semantic and metric information. Through the cross-frame alignment induced by the RPC and PSA modules, the framework enforces an implicit geometric consistency. As illustrated by the successful metric case in Fig. E(a), this mechanism allows the model to maintain stable object identities and consistent scale cues across multiple viewpoints with-

out the need for explicit 3D supervision. Consequently, EgoMind can more effectively leverage the implicit spatial priors inherent in MLLMs to support complex metric reasoning tasks, such as estimating room sizes or determining precise physical distances.

Failure Cases and Extensibility. Despite these strong spatial modeling capabilities, we identify two primary failure modes in highly complex scenarios. The first is *anchor mismatch* (Fig. Eb), which typically arises when environments contain multiple visually identical or similar objects, occasionally confusing the model's cross-frame object tracking. The second failure mode stems from *abrupt perspective shifts* (Fig. Ec), where severe or discontinuous camera movements lead to sparse visual anchors, breaking the coherent spatial narrative constructed by the RPC module.

Nevertheless, the linguistic nature of EgoMind renders it highly extensible. While fine-grained metric precision can be challenging for pure 2D MLLMs, incorporating partial metric hints (e.g., basic object size cues) into the prompt can significantly mitigate these issues. In our exploratory experiments, providing such hints improved the *Room Size*

estimation accuracy on VSI-Bench from 40.35% to 44.72%, demonstrating the flexibility and adaptability of our CoT framework.

D. Zero-Shot Performance

Remarkably, even for Gemini 2.5 Pro—a closed-source model—simply injecting EgoMind CoT as a prompting template (Fig. G) already elicits noticeably stronger spatial reasoning. As shown in Fig. F, the guided reasoning structure enables Gemini 2.5 Pro to consistently capture cross-frame correspondences, recognize implicit spatial bridges, and assemble a more coherent global scene representation. Quantitatively, applying this EgoMind CoT prompt yields a substantial zero-shot improvement on VSI-Bench for Gemini 2.5 Pro, boosting its overall accuracy from 50.62 to 59.73.

Conversely, applying the same zero-shot prompt to Qwen2.5-VL-7B yields only marginal gains from 30.02 to 32.89. This contrast reveals that while zero-shot CoT prompting alone can activate spatial reasoning in massive closed-source models, it is insufficient for smaller open-source models due to limited instruction-following capacities. Consequently, our two-stage training (SFT + RL) is indispensable for smaller models to fully internalize the reasoning paradigm, effectively driving performance to 50.16.

These findings demonstrate that EgoMind CoT is not bound to a specific model architecture. Instead, it serves as a generalizable and effective reasoning paradigm that substantially enhances spatial understanding in both open-source and closed-source MLLMs.